ILLINOIS

UNIVERSITY OF ILLINOIS AT URBANA-CHAMPAIGN

PRODUCTION NOTE

University of Illinois at
Urbana-Champaign Library
Large-scale Digitization Project, 2007.

Technical Report No. 447

VERBAL REPORTS OF THINKING AND MULTIPLE-CHOICE CRITICAL THINKING TEST DESIGN

Stephen P. Norris
Institute for Educational Research and Development
Memorial University of Newfoundland
and
Center for the Study of Reading
University of Illinois at Urbana-Champaign

January 1989

Center for the Study of Reading

TECHNICAL REPORTS

UNIVERSITY OF ILLINOIS AT URBANA-CHAMPAIGN
174 Children's Research Center
51 Gerty Drive
Champaign, Illinois 61820

CENTER FOR THE STUDY OF READING

A READING RESEARCH AND EDUCATION CENTER REPORT

Technical Report No. 447
VERBAL REPORTS OF THINKING AND
MULTIPLE-CHOICE CRITICAL THINKING
TEST DESIGN

Stephen P. Norris
Institute for Educational Research and Development
Memorial University of Newfoundland
and
Center for the Study of Reading
University of Illinois at Urbana-Champaign

January 1989

University of Illinois at Urbana-Champaign
51 Gerty Drive
Champaign, Illinois 61820

The work upon which this publication was based was supported by the Social Sciences and Humanities Research Council of Canada (Grant No. 418-81-0781). It was completed while the author was a Visiting Scholar at the Center for the Study of Reading. The views expressed herein are those of the author and not the funding agency.

Abstract

A methodology is described for using verbal reports of thinking to develop and validate multiple-choice tests of critical thinking. The procedure included the following steps: devising ideal models of thinking for each item; developing a method for interviewing subjects; collecting verbal reports of thinking from samples of subjects; using the ideal models to rate the quality of the thinking portrayed in the reports; comparing the quality of thinking to answer choices and accumulating those comparisons across subjects for each item; identifying and modifying suspect items. The methodology is offered as a way to gather evidence on the truth of claims that currently available multiple-choice critical thinking tests do not measure critical thinking (McPeck, 1981; Petrie, 1986; Whimbey, 1985) and as a way to develop new critical thinking tests.

VERBAL REPORTS OF THINKING AND MULTIPLE-CHOICE CRITICAL THINKING TEST DESIGN

During the last decade there has been a growing interest in teaching and assessing critical thinking. But there are few readily available instruments for measuring critical thinking. Furthermore, most of the ones that are available are multiple-choice and many advocates of critical thinking are particularly dissatisfied with multiple-choice testing formats (McPeck, 1981; Petrie, 1986). Ignoring for the moment the empirical support for this dissatisfaction, many critical thinking theorists insist that the development and validation of critical thinking tests require direct evidence on the thinking processes the tests require. However, most multiple-choice critical thinking tests are not supported by such evidence. It seems that in order for multiple-choice testing of critical thinking to be justified (or seen as justified), then design procedures must be implemented which somehow provide more direct evidence on the thinking processes that the tests elicit.

This paper describes how verbal reports of examinees' thinking on draft items of multiple-choice critical thinking tests can be used systematically to provide such evidence. Testing theorists have endorsed verbal reports of thinking for amplifying the meaning of the constructs a test is measuring (Cronbach, 1971, p. 474), for representing those constructs by the mental processes which underlie performance (Embretson, 1983), for directly analyzing the processes underlying item performance (Messick, 1987, p. 85), or for specifying the intellectual processes used to perform the task (Anastasi, 1988, p. 159). Such endorsements dovetail with the criticisms that there is no direct evidence that multiple-choice critical thinking tests in fact measure critical thinking (Ennis & Norris, in press; McPeck, 1981; Petrie, 1986; Whimbey, 1985).

There are very few reported uses of verbal reports of thinking in test development and validation (Norris, in press). When verbal reports have been used, the information in them rarely has been analyzed and compiled systematically. Rather, the information has been used casually to support intuitive judgments about whether items mislead, contain unwanted clues, use unknown vocabulary, and the like. These are important considerations in test design, but it would be useful to see how verbal reports of thinking might be used systematically to help make construct validity judgments about multiple-choice critical thinking tests. As an example, this paper reports how verbal reports of thinking were used to develop and validate a multiple-choice critical thinking test of observation appraisal. With suitable adjustments, the approach could be used to examine the validity of multiple-choice tests of other aspects of critical thinking.

Background

The Role of Observation in Critical Thinking

For the purposes of this study, Ennis's (1987) definition of critical thinking was adopted: "Critical thinking is reasonable and reflective thinking that is focused upon deciding what to believe or do." According to Ennis, thinking critically involves the use of basic information (information from others, from observation, and from previously drawn conclusions) and the process of inference to move from that basis to decisions about belief and action. Observations are central to critical thinking on this model, providing as they do some of the information upon which decisions are based. The test developed during this study is an attempt to measure the ability to appraise correctly reports of observations.

Motivation

Most definitions of critical thinking, including Ennis's, focus more on the *process* of thinking than its products. Paul (1985) argues that for learning to be rational "the process that leads to belief is more important than belief itself." It is a result of this focus on process that many critical thinking theorists

doubt that traditional multiple-choice tests can measure critical thinking, since such tests reveal products of thinking, not thinking itself (McPeck, 1981; Petrie, 1986).

This doubt is not pedantic. Criticisms of existing multiple-choice critical thinking tests have shown that keyed answers can be reached by thinking uncritically and unkeyed answers by thinking critically (Ennis & Norris, in press). The criticisms are usually hypothetical and not supported by empirical evidence on how examinees choose their answers. The point of the criticisms, however, is that there is no direct evidence one way or the other. Given the lack of evidence, there is a perceived unacceptable risk that some examinees are rewarded and some penalized unfairly. Such perceptions lead to a skepticism about the suitability of multiple-choice tests for measuring critical thinking.

A similar conclusion about multiple-choice critical thinking tests can be motivated by research on thinking in complex domains and within ill-structured problems. Such thinking often must follow multiple paths to multiple solutions (Resnick, 1987; Spiro, Vispoel, Schmitz, Samarapungavan, & Boerger, 1987). Thus, one might wonder how multiple-choice tests, which typically demand one solution, can test for critical thinking in complex domains.

There is, therefore, a concern among advocates of critical thinking that multiple-choice testing cannot serve well the goal of critical thinking evaluation. On the other hand, multiple-choice tests seem an appropriate format for testing certain aspects of critical thinking, such as knowledge of the many principles for appraising the credibility of information. So it is important to know whether they can be designed to provide more direct evidence on the thinking that underlies performance. Such direct evidence would place multiple-choice critical thinking tests on a sounder theoretical and empirical footing. Eliciting verbal reports of thinking on trial test items as described subsequently is one plausible way to collect the direct evidence desired.

Ability Test Validation

Construct validation should provide an explanation of performance on a test (Cronbach, 1971) by modelling the mental processes leading to the performance (Norris, 1983). According to this view, one criterion for judging construct validity of a critical thinking test is the extent to which critical thinking leads to good performance on the test and uncritical thinking to poor performance.

Construct validity must be inferred, not directly measured, but it is recognized that inferences about thinking processes based on item statistics and content analysis are untrustworthy (Connolly & Wantman, 1964, p. 59). Schuman (1966), for instance, has demonstrated that item performance can show excellent variation, correlate well with other item performances, and be related to other variables in relevant ways, even though the majority of examinees do not understand the item. As a result, some testing specialists recommend seeking more direct evidence on examinees' thinking than that contained in item statistics (Haney & Scott, 1987, p. 365), such as evidence from examinees' verbal reports of their thinking as they work through test items (Haney & Scott, 1987, pp. 301-2). Evidence of this sort can support inferences rarely supportable by item analysis data alone (Connolly & Wantman, 1964, p. 62).

Eliciting verbal reports of examinees' thinking is an attempt to gain information on the mental processes leading to their performance on the test, that is, to learn the information, mental strategies, and reasoning principles they use in choosing answers. The aim is to look beneath the surface of answer choices in order to gain more direct insight into the reasons for the choices, so that more trustworthy inferences can be made about the causes of performance and, derivatively, about construct validity. In the specific context of testing for critical thinking, the aim would be to use information in examinees' verbal reports of thinking to judge whether it is differences in critical thinking, and not something else, that accounts for differences in scores. In order for the procedure to work, a way to elicit verbal reports and of using the information contained in them needs to be designed and justified.

Past Use of Verbal Reports of Thinking in Test Design

Verbal reports of thinking have not been used often in test design. When such studies have been reported, there has been inadequate direction on how the information in verbal reports might be used to quantify the quality of examinees' thinking, on how to make comparisons between the quality of thinking as displayed in examinees' verbal reports and their choices of answers, and on how to accumulate and coordinate these comparisons across several examinees to help judge the validity of tests.

Bloom and Broder (1950) were among the earliest users of verbal reports of thinking in test development. They asked a sample of college students to think aloud while they worked through questions on a problem solving test. The verbal reports yielded relatively consistent and meaningful data from most students and helped to support inferences from the students' problem solutions to the thinking which led to the solutions.

Kropp (1956) used verbal reports of thinking to examine the relationship between choosing correct answers on 10 multiple-choice items and the reasoning used to make the choices. The items were designed to measure understanding of text with physical science content and were judged to involve more complex thinking than simply recalling information or locating it in the text. The verbal reports showed that a great variety of reasoning processes led to choosing correct answers. Since some of these processes were not justifiable, Kropp concluded that it is hazardous to infer the quality of reasoning processes on the basis of answers chosen. However, no procedure was described for determining the *tendency* for unjustified reasoning to lead to correct answers.

McGuire (1963) used verbal reports of thinking to make medical school certification examinations test for critical thinking and problem solving and not merely factual recall. Experts in the field were asked to classify test questions according to a modified Bloom's taxonomy. Medical students were then asked to think aloud as they worked through the items. McGuire reported that comparing the experts' classifications to the students' verbal reports helped her to align better the objectives of instruction and their assessment. However, it is unclear exactly how the verbal reports were examined and how the information in them was used to develop the test. As in the previously described studies, there is no specific guidance on how verbal reports of thinking can be used in test development.

Connolly and Wantman (1964) reported a study intended to improve the one of Bloom and Broder. They studied nine subjects, who were asked to report all the thoughts that might cross their minds as they worked on a set of 25 verbal analogy items. The adequacy of the subjects' thinking was compared to a model of ideal thinking on the items. The model was general, rather than item specific, and listed 67 behaviors representing quality of thinking (e.g., "Justifies choice with incorrect logic," and "Reads stem and immediately states keyed response"). The attempt to use an ideal model of reasoning to judge verbal reports is significant, but insufficient detail is provided to know what the model is, where it came from, and how it was used. They concluded that their technique was useful for pretesting items, but did not mention a role for the technique in test validation. They suggested that the method might be used to determine whether some students are penalized for having too much knowledge. This suggestion is particularly relevant to multiple-choice critical thinking testing, because of the allegations that it is knowledge and not critical thinking that these tests examine.

A study by Schuman (1966) provides the most detailed procedure for using the information in verbal reports of examinee's thinking. One thousand factory workers and cultivators in East Pakistan completed an attitude survey and then reported on the reasons for their responses to 10 randomly selected items. Their reasons were rated on a scale from "1," given to a reason that was very clear and led to an accurate prediction of the chosen response, to "5," given to a reason that was very unclear and could not be used to predict the answer chosen. Average scores were calculated for each item across all individuals and were used to indicate the subjects' understanding of that item; the lower the score the better the understanding.

Schuman's procedure for quantifying and using the information in examinees' verbal reports marks an important difference from previous studies. But the approach is not without problems. For instance, examinees' reasoning could clearly express a misunderstanding of an item, yet reliably lead to predictions of their responses. Their reasoning would be assigned "1," and the item would receive a low average score, even though subjects did not understand it.

In sum, the research on using verbal reports of thinking for developing and validating tests is not adequate. There have been few systematic attempts to use the information contained in verbal reports to quantify the quality of examinees' thinking and to accumulate the ratings of quality for each item across individuals. Furthermore, little emphasis has been placed on developing models against which reports of thinking might be compared. Verbal reporting techniques have been used rather casually in designing multiple-choice tests, with the possible exception of Schuman's study.

A Method for Eliciting and Using Verbal Reports of Thinking

To systematically use verbal reports of thinking in the development and validation of multiple-choice critical thinking tests, particular steps should be followed. This section will use the specific example of developing and validating the Test on Appraising Observations (Norris & King, 1983) to illustrate the need for the following general steps:

1. developing a set of critical thinking principles which items would serve to test;
2. devising initial trial items to test the principles;
3. developing models of ideal thinking on each item;
4. developing an interviewing methodology for eliciting trustworthy reports of thinking;
5. collecting verbal reports of thinking on items;
6. scoring verbal reports using the models of ideal thinking as a standard and scoring answers chosen according to the answer key;
7. comparing thinking on items to performance on them and systematically accumulating these comparisons across examinees for each item;
8. modifying suspected items;
9. retrying the test in accord with steps 5 to 8.

Developing Principles of Observation Appraisal

In order to assess critical thinking and to compare the quality of examinees' verbal reports of thinking to their choice of answers, there needs to be some justified standard for rating quality of thinking. The Test on Appraising Observations was designed to assess ability to appraise reports of observations, so an early task was to develop a comprehensive and defensible set of principles for appraising observations. Table 1 presents a sample of the principles which are described comprehensively and defended in Norris (1984) and Norris and King (1984).

[Insert Table 1 about here.]

The principles do not provide either necessary or sufficient conditions for observation statements to be credible. They suggest conditions to consider. To the extent that these conditions are satisfied, an

observation is worthy of belief, and to the extent they are not satisfied, it is worthy of doubt. In any given case, principles may compete. There are no strict procedures for arbitrating such difficult cases; good judgement and experience are needed.

The principles find support from a number of sources. For example, those dealing with conflict of interest, leading questions, and the skill of the observer find support in judicial practice (Brooks, 1983). The principle about leading questions is also supported by psychological research on eyewitness testimony (Lindsay, Wells, & Rumpel, 1981; Loftus, 1979; Wells, Ferguson, & Lindsay, 1981; Yarmey, 1979). Several principles are supported by common-sense psychology. For example, it makes common sense to think that an observation is more believable when made by a more alert observer.

Devising Initial Trial Items

Initial trial items were written to satisfy a number of criteria. They were in multiple-choice format and each was written to test ability to use one principle of observation appraisal. Fifty items were cast in the context of two stories: Part A, the story of a traffic accident, and Part B, the story of a river exploration. The story context was used in order to provide examinees sufficient context for deciding which principle of observation appraisal should be applied.

Each item presented two conflicting reports of observations made by characters in the stories and required examinees to decide which, if either, of the reports is more credible. Comparative judgments were requested because without the wealth of information available in real situations it is usually not possible to make absolute judgments about the credibility of single observation statements (Ennis, 1988; Norris & Ennis, in press).

Thus, for each item the three possible answer choices were: the first statement is more credible, the second statement is more credible, and the statements are equally credible. Multiple-choice items usually have more than three alternatives to minimize problems of guessing, but only the above three were used because they are the logically obvious ones given the task.

Developing Models of Ideal Thinking on Each Item

The principles of observation appraisal were used to develop a model of ideal thinking on each item. As an example, consider Item 3 from Part A, the traffic accident story:

> A policewoman has been asking Mr. Wang and Ms. Vernon questions. She asks Mr. Wang, who was one of the people involved in the accident, whether he had used his signal.
> Mr. Wang answers, "Yes, *I did use my signal.*"
> Ms. Vernon had been driving a car which was not involved in the accident. She tells the officer, "*Mr. Wang did not use his signal.* But this didn't cause the accident."

Examinees were to choose which, if either, of the italicized statements is more credible. To reason through this question correctly, an examinee first needs to be able to derive from the text the relevant information about Wang's and Vernon's involvement. The text is simple enough that most high school students should have no difficulty with this task. Second, an examinee must retrieve from background knowledge the relevant facts that not using a turn signal can cause an accident and that being held responsible for an accident can be troublesome. Again, high school students would have such common knowledge. Finally, an examinee has to infer that the possibility that one's own testimony can place one in trouble is an accuracy-reducing factor and, thus, that Wang is less credible than Vernon.

> The ideal model of thinking developed for this item is as follows: Mr. Wang was involved in the accident, but Ms. Vernon was not involved. Mr. Wang is less credible because his involvement would give him reason to say he used his signal even if he did

not. Wang is in a conflict of interest. People in a conflict of interest, that is, people who have something to gain by events being cast as they described them, tend to be less credible than those who are not in such a situation.

The reasoning required for Item 3 generalizes to the other test items. In order to reason ideally through the items examinees need to do the following, though not necessarily in the order indicated:

1. Cite the relevant facts in the text which can be used to compare the credibility of the underlined statements;

2. Use these facts together with relevant background knowledge to make a comparative evaluation of the credibility of the statements;

3. Show how the evaluation is based on an appropriate general principle which subsumes the relevant facts in the text under a general accuracy-reducing factor.

Each item on the test was analyzed according to this scheme, models of ideal reasoning for each item developed, and the models used in a manner described subsequently to rate the quality of examinees' verbal reports of thinking.

Developing an Interviewing Methodology

An interviewing methodology was designed to elicit examinees' reasoning while they worked through items on the observation test. In accord with widely advocated practices for collecting verbal reports as data (Afflerbach & Johnston, 1984; Ericsson & Simon, 1980, 1984; Larkin & Rainard, 1984; Loftus, 1979; Norris, in press) the interviews were as non-leading as possible. The initial directive to examinees was: "As you do each question tell me all you can about what you are thinking while you are picking your answer." At this stage, interrupting an examinee's verbal reporting was permitted only to clarify the ambiguous referent of a pronoun or to point to obvious reading errors.

It is important to obtain records of reasoning that are as complete as possible. To fulfill this aim it is necessary sometimes to probe beyond the initial instruction to think aloud, while being careful not to rush examinees by prematurely cutting off their reasoning (Larkin & Rainard, 1984, p. 250), or to endorse or criticize particular reasoning attempts. Even in such follow-up stages, it is desirable for probing to be as non-leading as possible, by merely echoing examinees' reported thoughts or by asking them to explain a little more what they have said.

To meet the above goals and constraints, the interview model given in Appendix A was adopted. It has three stages. In Stage One the interviewer informs examinees of the general purpose of the interviews and of their role in them, especially the fact that they will be asked some questions to be answered verbally.

In Stage Two the interviewer asks examinees to tell all they can about what they are thinking while choosing their answers. Interruptions to examinees' reports are permitted only to probe for ambiguous references and for obvious reading errors (See II.2.1 and II.2.2). If examinees ask for additional information or for reasons, the interviewer must respond as in II.3.1, "You can only go by what is written," or II.3.2, "You can decide only according to what is said and what you know." If examinees seek feedback on their progress it must not be given (II.4.3).

Stage Three begins after examinees have chosen their response and have finished reporting on their thinking in response to the directive in (II.1). The interviewing model for this stage consists of a set of condition-action pairs which guide the actions of the interviewer in accord with the satisfaction of conditions defined by a combination of examinees' answer choices and the content of their verbal reports. Each action is designed to match a particular combination by being as non-leading as possible

and still having a chance to elicit useful information. For any given item and examinee response, only one numbered step will apply. The strategies listed under III.4 guide the interviewer when it is not clear how to proceed given an examinee's report.

The model for Stage Three is used as follows. Starting with III.1 and proceeding sequentially, the interviewer checks whether the condition (described by the phrase following the "if") is satisfied. Sometimes the condition is complex, that is, consists of subconditions, so all subconditions must be met for the condition to be satisfied. If a condition is not satisfied, the interviewer proceeds to the next numbered step. If satisfied, the interviewer carries out the indicated action (described by the phrase following the "then"). When the signal, "*then proceed to the next item*," is reached the interview reverts to Stage Two.

Collecting Verbal Reports of Thinking

Our experience has shown that a multiple-choice test takes about four-times as long in a verbal reporting format as in a paper-and-pencil format. Therefore, it is too tiresome for students to be interviewed on a whole test designed for a regular class period in paper-and-pencil format. However, if there are story lines in a test, as is the case with the Test on Appraising Observations, or the items are in a logical sequence for another reason, a student could not be interviewed on a later section without having completed previous sections. A procedure for interlacing interviewing and paper-and-pencil formats is needed.

In the example study there were two interviewers. One-half the students were assigned randomly to each. Within these halves, half was assigned to take Part A of the test and the other half to take Part B. Each part of the test was divided into two sections representing logical breaks in the story line: Part A, Section I (Items 1-15) and Section II (Items 16-28); Part B, Section I (Items 29-37) and Section II (Items 38-50). One-half of the subjects assigned to a part were randomly chosen to be interviewed on Section I of that part and to complete Section II in paper-and-pencil format; the other half was to write Section I and be interviewed on Section II.

Each interviewer dealt with two students at a time working on the same part, either A or B, of the test. First, both students were told how the testing would proceed. The purpose of the study had already been explained in meetings with students in their classes. The student who was to write Section I of the part was shown to a place to work alone, told to stop when finished the last item of Section I (either Item 15 or 37), told not to read on, and to wait to be called by the interviewer. The other student was interviewed on Section I according to the model in Appendix A. When the interview was over, the students exchanged places. The student who wrote Section I was interviewed on Section II and the student who was interviewed on Section I wrote Section II. Students marked their answers to all questions on a standardized answer sheet.

Scoring Verbal Reports of Thinking and Answer Choices

Performance scores were assigned to answer choices for each item according to whether they agreed (1) or disagreed (0) with a key based on the set of principles illustrated in Table 1.

Verbal reports were transcribed verbatim. Each student was assigned a *thinking score* for each item according to the rating scale in Table 2. The rating scale is based upon the ideal models of thinking and assigns scores according to how closely students' verbal reports approximate the models. Table 3 shows how the rating scale would work for Item 3.

[Insert Tables 2 & 3 about here.]

For illustration, here are the verbatim transcriptions of the verbal reports of two senior high school students working on Item 3:

Student A . . .ah . . .ah, Mr. Wang, like he probably didn't, like you know, it was just, he probably thought he used his signal, but really didn't. And Miss Vernon, she was watching so she'd be able to tell from back if he was using it or not. Right? Being the case, so, I'd tend to believe Vernon.

Student B I would say that he did use his signal because anybody who's in the car . . . coming up to an intersection or anything . . . he, he usually knows what he's doing. So I'd be more inclined to believe the first.

Student A chose the keyed response and Student B an unkeyed one. Neither student thought critically on the item, however, and both were assigned zero for thinking scores. The criteria used by both students for comparing Wang and Vernon were unsound. Student A's reasoning that Wang just "thought he used his signal," but Vernon would "be able to tell from back" is arbitrary. It is used to rationalize a choice of answer rather than to justify it. It is just as reasonable to say that Wang would "be able to tell from inside that he used his signal" and that Vernon just "thought he did not use his signal." Student B's reasoning fails to distinguish between Wang's *knowing* what he had done and his *telling accurately* what he had done, and to allow that someone in another car can know what another driver is doing. So the students' performance scores and thinking scores for Item 3 are as in Table 4.

[Insert Table 4 about here.]

For each student, there was a performance score for each item taken and, for each item done in the interview, a thinking score also. Performance and thinking scores were conceptually and empirically independent, as illustrated by the fact that both Student A and Student B thought poorly, but one answered correctly Item 3 and the other did not. Thinking scores were assigned without reference to the answer chosen. When appropriate, students were given credit for using correctly criteria other than those assumed when designing the items. That is, the ideal model of thinking for each item was used as a guide, but it was possible for reasoning to be rated highly according to general standards of critical thinking even though it was outside the bounds of the model. For instance, an examinee might reject a factual presupposition built into an item and reason perfectly acceptably from an alternative presupposition to other than the keyed response. In such a situation, the thinking score would be high, but the performance score would be "0."

Comparing Performance and Thinking Scores and Judging Items

A central feature of the test development and validation methodology being described is the comparison of performance and thinking scores in the evolution of a test. The aim is to adjust items, or to write new ones, to make a strong relationship between quality of thinking and quality of performance for each item. The relationship was calculated in two ways: using biserial correlation coefficients and a thinking/performance index (t/p index) which estimates the net evidence, positive minus negative, provided for an item by the combinations of examinees' performance and thinking scores.

Performance scores constitute a dichotomous variable and thinking scores an ordinal variable, both with assumed underlying normal distributions. The biserial correlation (r_{bis}) is the best to use for relating variables of this sort, but two problems exist. First, when one of the variables has zero variance, (e.g., when all examinees get an item wrong), then $r_{bis} = 0$ for the item regardless of scores on the other variable. However, when all of the ordered pairs of thinking scores and performance scores for an item are (0,0), then there is a perfect coincidence between thinking poorly and getting the item wrong. This coincidence is not reflected in the $r_{bis} = 0$.

The second problem is that when certain assumptions underlying the statistic are violated, $r_{bis} > 1$ (Kendall & Stuart, 1961). The assumption of underlying normality is crucial for this statistic. In

addition, r_{bis} is a most efficient estimator of the population parameter (r_{pop}) when the dichotomy between getting an item right and wrong lies in the middle of the underlying distribution of ability on the item, and when $r_{pop} = 0$. When $r_{pop} = - > 1$, r_{bis} is least efficient. For $n < 15$, which was the case in this study, $r_{bis} > 1.25$ is common.

The thinking/performance index was developed in order to avoid these problems with using the biserial correlation and also to capture more intuitively the sense of positive and negative evidence for items contained in examinees' performance and thinking scores. The t/p index is an average, across all subjects thinking aloud on an item, of the evidence for the quality of that item contained in the relationships between each subject's thinking and performance scores.

To calculate the t/p index for an item, combinations of thinking and performance scores for each subject first were weighted as in Table 5. Thinking scores from 0 to 2 only are represented in the table, since so few scores of 3 were obtained in the samples studied. Any thinking score of 3 was converted to 2. There are thus six possible combinations of thinking scores and performance scores. The weightings assign ordered degrees of evidence for the quality of an item provided by each of these combinations.

[Insert Table 5 about here.]

Thinking-score/performance-score combinations (0,0) and (2,1) were judged to give the same degree of positive evidence for the quality of an item, a weighting of +2. The evidence is positive because, in the first case, poor thinking is associated with choosing an unkeyed answer and, in the second case, good thinking is associated with choosing the keyed response. The weighting of +2 was assigned to reflect that these combinations provide the greatest positive evidence for an item that can be obtained.

Combinations (0,1) and (1,0) were judged to provide the same degree of negative evidence for item quality, a weighting of -1. The evidence is negative in the first combination because the person would have thought poorly, but chosen the keyed answer nevertheless. In the second combination, the person would have thought well but chosen an unkeyed answer. Combination (2,0), assigned a weighting of -2, was judged to provide more negative evidence than either of the previous two combinations, since the subject would have thought *very* well but selected an unkeyed response. Finally, combination (1,1) was judged to provide positive evidence for an item, but not as positive as the combination (2,1), so was assigned a weighting of +1.

To compute the t/p index for an item, the weightings of evidence for each subject answering the item are assigned according to Table 5, averaged, then divided by 2. The index thus defines a range of evidential weight from -1 to +1. For students A and B on Item 3 (see Table 4), the weights of evidence would be -1 and +2, respectively. The t/p index would be .25. This t/p index is *not* a measure of the correlation between thinking and performance scores.

The t/p index captures intuitions about the meaning of the data which r_{bis} cannot. Consider the following two sets of ordered pairs of thinking and performance scores for Items 18 and 38 in Version B of the Test on Appraising Observations:

{18:(2,1) (0,0) (0,0) (2,1) (2,1) (2,1) (2,1) (0,0) (2,1) (2,1) (1,1) (2,1) (2,1) (2,1)}; {38: (0,1) (0,0) (0,1) (0,0) (0,0) (0,0) (1,0) (0,0) (0,1) (0,0) (1,0) (1,1)}.

For Item 18 all of the thinking/performance combinations are in a positive direction. That is, for each subject, choosing the keyed answer is associated with thinking well and choosing an unkeyed answer is associated with thinking poorly. This relationship is reflected in a high $r_{bis} = 1.338$, but of course it is not clear what such a correlation means. However, comparing Item 38 for which there are 5 mismatches between thinking and performance scores (either combinations (1,0) or (0,1)), the computed r_{bis} is still high (1.296). Thus, the computed r_{bis} does not agree with the intuition that for Item 18 there is a much higher correspondence between thinking well and performing well than for

Item 38. However, the comparison of t/p indexes for the items does agree with intuition: for Item 18, the t/p index = .964; for Item 38, it is .334.

I cannot be very specific about the meaning of different sizes of the t/p index, because its properties need further exploration. But some general things can be said. First, a negative t/p index indicates that there is more evidence against the quality of an item than there is for it. Such an item should be discarded or modified. An approximate lower bound of acceptability can be derived on the assumption that a t/p index should be at least as high as the index which would obtain for random guessing. In such a situation, all examinees would receive 0 for thinking scores, since quessing randomly is not thinking critically. For a sample of N students working on the Test on Appraising Observations, N/3 would on average choose the keyed response, receive thinking/performance score combinations of (0,1) and evidence weightings of -1. The remaining 2N/3 students would choose an unkeyed response, receive score combinations of (0,0) and evidence weightings of +2. The t/p index would be .5. So, .5 is a reasonable approximate lower bound of acceptability for t/p indexes on a three-alternative multiple-choice test. For a four-alternative test, the lower bound would be .625. Note that these are conservative lower bounds, since in the situation of random guessing two-thirds of the sample in the three-alternative case and three-fourths in the four-alternative case receives the highest possible evidence rating of +2.

As a sample set of data, Table 6 contains t/p indexes and biserial correlations between item and test performance (not biserials between thinking and performance scores as have been discussed up to this point) for Versions B and C of the Test on Appraising Observations. The t/p indexes and item/test biserials are not highly correlated: .12 ($p = .42$) for Version B and .30 ($p = .03$) for Version C. Therefore, the t/p indexes give different information on item quality than the biserials, so decisions made about items using the t/p indexes would likely differ from decisions made using the biserials. This fact suggests that the t/p indexes are worth further exploration, because at least in this research they did not duplicate information derived from traditional item statistics.

[Insert Table 6 about here.]

Modifying Suspected Items

Item level statistics, such as item difficulty levels, item/test biserial correlations, and t/p indexes, were used to identify problematic items in experimental versions of the test, but most weight was given to the t/p indexes. Rounded to one decimal place, there were 16 items with t/p indexes < .5, the computed lower bound of acceptability. Reasons were sought for these low indexes by examining the transcribed interviews and asking the following questions:

1. Do examinees understand the task in the intended way, or are the instructions unclear, is the item systematically misleading, or is the assumed background knowledge outside their knowledge store?

2. Are examinees thinking critically to arrive at other than the keyed response and, if so, should the keyed response or the item be changed?

3. Are examinees thinking uncritically yet arriving at the keyed answer and, if so, are there clues in the text that can be eliminated?

For Version B, all but two of the 16 suspect items, Items 22 and 23, were altered on the basis of information contained in the verbal reports. Alterations included single word changes, the addition and deletion of larger amounts of information, and changes to the directions. Items 22 and 23 were not altered, because they were close to the .5 standard of acceptability and it was expected that changes to the neighboring and thematically closely linked Items 20 and 21 would improve them also.

An example of changes made to the directions will serve to illustrate how the t/p indexes and the verbal reports were used together. Recall that Part A of the test was set in the context of a traffic accident. The Version B directions listed the names of all characters and what they were doing at the time of the accident. Since there were several characters in the story with different roles, this list should have helped to keep them straight. But the verbal reports showed that many examinees looked for direct answers to questions in this information. In one item, for example, characters gave conflicting reports about how many cars were at the intersection when the accident occurred. One character was more alert to the traffic and hence a more credible source of information about the number of cars at the intersection. Coincidentally, that character reported seeing the same number of cars *at the intersection* (three) as the directions mentioned were *involved in the accident*. The verbal reports showed that several examinees did not consider the characters' alertness, but merely counted the number of cars mentioned in the directions and chose the keyed answer as a result. These students thought uncritically, because they equated the number of cars at the intersection with the number involved in the accident, but chose the keyed response. It is the possibility of just this sort of problem with multiple-choice tests that advocates of critical thinking find most disturbing.

Retrying the Modified Test

When the 16 suspected items were revised, the new version, Version C, was tried by collecting verbal reports of thinking on it, calculating t/p indexes, and so on. The t/p indexes and biserial correlations between item and whole test performance for Version C are in Table 6.

The average t/p index for Version B was 0.58 and for Version C was 0.66, indicating that the changes had increased the quality of the test. There were no negative t/p indexes for Version C and, of the 16 items on Version B with indexes < .5, 10 had their indexes raised to > .5. Only one item of the 16 received a lower t/p index for Version C.

On the other hand, 5 items on Version C obtained t/p indexes < .5 which had indexes > .5 on Version B. The interviews for both versions were examined. For items 3, 4, 7, and 8 there was a common error in students' thinking which was largely responsible for the low t/p indexes. In each item, several students made their choice solely on the basis of who was the driver of a car, on the grounds that a driver would be more alert to the traffic than a pedestrian or passenger. For example, in Item 3 on Version C, Mr. Wang was a driver and Ms. Vernon "watched the accident happen." Several students chose Mr. Wang's statement as more believable because he was a driver and thus more alert to road conditions. These students were assigned thinking scores of 1 for recognizing this important factor, even though they neglected the more important fact that Wang was in a conflict of interest. Thus, they chose an unkeyed response but thought fairly well, contributing to a lowered t/p index. The obvious change was to make both Wang and Vernon drivers, as is the case in the version of Item 3 cited previously as an example.

The patterns of thinking and performance score combinations for the remaining seven items with t/p indexes < .5 were examined. No systematic problems could be discerned in the items, so it was decided not to make further changes.

Discussion and Conclusions

Verbal reports of thinking on trial items of multiple-choice critical thinking tests seem useful and important from at least three perspectives. First, existing multiple-choice critical thinking tests are often criticized because they give no direct evidence of the process of thinking that leads to examinees' choices of answers. If we assume that such direct evidence is needed, then the criticism is certainly justified, because multiple-choice critical thinking tests have traditionally been constructed without direct evidence on the thinking processes used to answer items on them.

Second, multiple-choice critical thinking tests could serve several useful purposes if they were valid. They are useful when testing groups of students, when time for grading is at a premium, and when knowledge of several criteria and principles of critical thinking is being tested. That is, multiple-choice critical thinking tests can fill a particular set of roles, though they likely cannot serve all the needs of critical thinking assessment. For instance, multiple-choice tests probably cannot assess ability to think critically in complex situations where solutions are radically underdetermined by available information.

Third, direct evidence on the information, strategies, and principles that examinees use to answer multiple-choice critical thinking items would be directly relevant to the construct validity of the tests. Construct validity is concerned with the causes of test performance. A critical thinking test would be valid to the extent that the cause of good performance is thinking critically and the cause of poor performance is thinking uncritically. Verbal reports of examinees' thinking on trial versions would provide information directly relevant to making such causal explanations of performance in the context of test use. In the context of test design, direct evidence from examinees' verbal reports would be gained on the causes of their performance. Based on this evidence, items would be discarded, modified, or retained so that, to the extent possible, the cause of performance is degree of the relevant aspects of critical thinking ability. In the context of use, the test would inherit the weight of the direct evidence gathered in the design context, even though in the use context no direct evidence would be gained on the causes of performance. If the examinees are reasonably similar in relevant characteristics to those sampled during the design, then there is a very small chance that for any given examinee there would be a systematic mismatch across items between the examinee's critical thinking and the answers chosen. Thus, by collecting direct evidence on thinking during the design stage of multiple-choice critical thinking tests, the advantages of relatively easy administration and scoring can be had while avoiding one of the main shortcomings of multiple-choice tests, their failure to provide direct evidence on thinking processes.

A verbal reporting methodology developed for the design of one test is not likely to be directly usable for another. So the procedure described in this paper is best thought of as exemplifying a methodology rather than providing a formula for other tests. The methodology has been adapted for the construction and validation of an inference test in reading comprehension for middle school students (Phillips, in press). In that study, in addition to using verbal reports as a source of evidence on the quality of items, verbal reports were used to construct distractors plausible to examinees who did not know the correct answers. The distractors were constructed to portray typical poor reasons which were given for answer choices. It would be worthwhile to study this technique with tests in other subjects.

In addition to test development, the verbal reporting procedure would be useful for evaluating the claims that currently available multiple-choice critical thinking tests do not measure critical thinking (Ennis & Norris, in press; McPeck, 1981; Petrie, 1986; Whimbey, 1985). For instance, the procedure could be adpated for studying the validity of the Watson-Glaser Critical Thinking Appraisal (Watson & Glaser, 1980) and the Cornell Critical Thinking Test Level X (Ennis & Millman, 1985), two of the oldest and most widely used multiple-choice critical thinking tests.

There are a number of possible outcomes of such studies. First, there is the possibility that the criticisms of these tests would be sustained. This would be unfortunate, because it would cast doubt on much of the research on the effectiveness of critical thinking instruction which used the tests as criterion measures. Another possible result is a clearer understanding of the range of applicability of existing critical thinking tests. Many multiple-choice critical thinking tests are advertised for use over wide age ranges. For example, The Cornell Critical Thinking Test Level X is said to be usable from Grade 4 to Grade 14 and the Watson-Glaser Critical Thinking Appraisal from Grade 9 on up. These ranges include individuals who differ widely in intellectual sophistication, empirical knowledge of the world, and test-taking experience, so it is quite plausible that the tests are not equally suitable across the entire ranges. A study of the verbal reports of thinking of examinees from different groups would provide direct evidence on the range of suitability of each test. A possible consequence is that the

range for each test will need to be narrowed. Another possibility is that answer keys might be adjustable to match more suitably the critical thinking of examinees of different age levels.

The verbal reporting approach itself needs further study. One aspect needing examination is the procedure for comparing thinking and performance scores. The biserial correlation is clearly not suitable, but the t/p index needs further analysis. In particular, a closer analysis of the meaning of the t/p index is required, especially of the minimally acceptable level of the t/p index to declare that an item is working adequately. In this study, the t/p index for random responses was taken as the minimally acceptable value. But several items on the final version of the Test on Appraising Observations received lower indexes than this, even though it was not obvious how the items might be improved.

The elicitation of verbal reports of thinking also needs to be studied with respect to different types of test content, different elicitation procedures, and different types of examinees. It is known, for instance, that different questioning procedures can affect differently the reports eyewitnesses give of events (Loftus, 1979). Can different approaches to eliciting verbal reports of thinking alter what examinees report? One study (Norris, in press) shows that verbal reports of high school students' thinking are not easily altered by asking them more leading questions, and that performance scores while giving verbal reports are the same as performance scores in a paper-and-pencil test taking situation. This is promising. But would verbal reporting work as well with other groups and other testing situations?

Indirect evidence for the value of the approach can be found by examining standard indicators of test validity and reliability. The Cornell Critical Thinking Test (Ennis & Millman, 1985) and the Watson-Glaser Critical Thinking Appraisal (Watson & Glaser, 1980) provide criteria for judging the validity of the Test on Appraising Observations and hence of indirectly judging the effectiveness of the verbal reporting technique. It was expected that the Test on Appraising Observations should be about as reliable as the Cornell and Watson-Glaser Tests. Those tests have, respectively, 71 and 80 items, whereas the Observations Test has 50 items. This fact would tend to make the Observations Test less reliable. However, the Cornell and Watson-Glaser Tests both attempt to measure several aspects of critical thinking, while the Observations Test focusses on one aspect, facts which would tend to make items on the Observations test intercorrelate more highly and its reliability higher. On balance, then, the reliability of the Test on Appraising Observations was expected to be about equal to the other two Tests.

Table 7 reports the range of reliabilities for the three tests computed from the performance of four groups of senior high school students from Southern Ontario, Canada. The mean reliability for the Observations test is less than that for the other two tests, though approximately equal that for the Cornell Test. If the outliers from the Test on Appraising Observations (Group D) and from the Watson-Glaser Test (Group A) are dropped, the mean reliabilities for the three tests, are virtually identical.

[Insert Table 7 about here.]

It was also expected that the correlation of the Observations Test with the Cornell Test should be higher than with the Watson-Glaser Test (assuming the three tests to be valid). The rationale for this expectation is that the Cornell Test includes a section of 24 items on judging the credibility of reports of observations, but the Watson-Glaser has no items on judging credibility.

Table 8 provides correlations among the three tests for the four groups of Ontario students. In three of the four cases the correlation with the Cornell Test is higher than with the Watson-Glaser. But the correlations are not too high, indicating that the Observations Test is measuring something different from the other two tests.

[Insert Table 8 about here.]

The verbal reporting methodology thus led to a test which meets several expectations held for it. This is evidence for the value of the methodology. Therefore, eliciting verbal reports of examinees' thinking on multiple-choice critical thinking tests, comparing the information in the verbal reports to examinees' choices of answers, and using these comparisons to systematically weigh and balance the evidence for the quality of items, seems a valuable approach for developing and validating such tests. In particular, it is a way to satisfy two competing desires of many proponents of critical thinking: the desire to use multiple-choice critical thinking tests, because they provide a convenient way to gather information on students' critical thinking, and the desire to have information on the process, not just the products, of that thought.

References

Afflerbach, P., & Johnston, P. (1984). On the use of verbal reports in reading research. *Journal of Reading Behavior, 41*(4), 307-322.

Anastasi, A. (1988). *Psychological testing.* New York: Macmillan.

Bloom, B. S., & Broder, J. L. (1950). *Problem-solving processes of college students.* Chicago: The University of Chicago Press.

Brooks, N. (1983). *Police guidelines: pretrial eyewitness identification.* Ottawa: Law Reform Commission of Canada.

Connolly, J. A., & Wantman, M. J. (1964). An exploration of oral reasoning processes in responding to objective test items. *Journal of Educational Measurement, 1,* 59-64.

Cronbach, L. J. (1971). Test validation. In R. L. Thorndike (Ed.), *Educational measurement* (2nd ed., pp. 443-507). Washington, D.C.: American Council on Education.

Embretson (Whitely), S. (1983). Construct validity: Construct representation versus nomothetic span. *Psychological Bulletin, 93,* 179-197.

Ennis, R. H. (1987). A taxonomy of critical thinking dispositions and abilities. In J. B. Baron & R. Sternberg (Eds.), *Teaching thinking skills: Theory and practice* (pp. 9-26). New York: W. H. Freeman.

Ennis, R. H. (1988). Testing teachers' competence, including their critical thinking ability. In B. Arnstine & D. G. Arnstine (Eds.), *Philosophy of Education 1987* (pp. 413-420). Normal, IL: Philosophy of Education Society.

Ennis, R. H., & Millman, J. (1985). *Cornell critical thinking test, level X.* Pacific Grove, CA: Midwest Publications.

Ennis, R. H., & Norris, S. P. (in press). Critical thinking evaluation: Status, issues, needs. In J. Algina & S. M. Legg (Eds.), *Cognitive assessment of language and mathematics outcomes.* Norwood, NJ: Ablex.

Ericsson, K. A., & Simon, H. A. (1980). Verbal reports as data. *Psychological Review, 87,* 215-251.

Ericsson, K. A., & Simon, H. A. (1984). *Protocol analysis: Verbal reports as data.* Cambridge, MA: MIT Press.

Haney, W., & Scott, L. (1987). Talking with children about tests: An exploratory study of test item ambiguity. In R. O. Freedle & R. P. Duran (Eds.), *Cognitive and linguistic analyses of test performance* (pp. 298-368). Norwood, NJ: Ablex.

Kendall, M. G., & Stuart, A. (1961). *The advanced theory of statistics.* London: Griffin.

Kropp, R. P. (1956). The relationship between process and correct item responses. *Journal of Educational Research, 49,* 385-388.

Larkin, J. H., & Rainard, B. (1984). A research methodology for studying how people think. *Journal of Research in Science Teaching, 21*(3), 235-254.

Lindsay, R. C. L., Wells, G. L., & Rumpel, C. M. (1981). Can people detect eyewitness-identification accuracy within and across situations? *Journal of Applied Psychology, 66*, 79-89.

Loftus, E. F. (1979). *Eyewitness testimony*. Cambridge, MA: Harvard University Press.

McGuire, C. (1963). Research in the process approach to the construction and analysis of medical examinations. *National Council on Measurement in Education Yearbook, 20*, 7-16.

McPeck, J. (1981). *Critical thinking and education*. New York: St. Martin's Press.

Messick, S. (1987). *Validity*. Princeton, NJ: Educational Testing Service.

Norris, S. P. (1983). The inconsistencies at the foundations of construct validation theory. In E. R. House (Eds.), *Philosophy of evaluation* (pp. 53-74). New Directions for Program Evaluation, No. 19. San Francisco: Jossey-Bass.

Norris, S. P. (1984). Trying eyewitness testimony, validating tests, and teaching critical thinking. In R. E. Roemer (Ed.), *Philosophy of education 1983* (pp. 179-189). Normal, IL: The Philosophy of Education Society.

Norris, S. P. (in press). Informal reasoning assessment: Using verbal reports of thinking to improve multiple-choice test validity. In D. N. Perkins, J. Segal & J. F. Voss (Eds.), *Informal reasoning and education*. Hillsdale, NJ: Erlbaum.

Norris, S. P., & Ennis, R. H. (in press). *Evaluating critical thinking*. Pacific Grove, CA: Midwest.

Norris, S. P., & King, R. (1983). *Test on appraising observations*. St. John's, Newfoundland: Institute for Educational Research and Development, Memorial University of Newfoundland.

Norris, S. P., & King, R. (1984). Observation ability: Determining and extending its presence. *Informal Logic, 6*(3), 3-9.

Paul, R. (1985). Bloom's taxonomy and critical thinking instruction. *CT News, 3*(5), 3-6.

Petrie, H. (1986). Testing for critical thinking. In D. Nyberg (Ed.), *Philosophy of education 1985* (pp. 3-19). Normal, IL: Philosophy of Education Society.

Phillips, L. M. (in press). *The design and development of the Phillips-Patterson test of inference ability in reading comprehension*. Champaign: University of Illinois, Center for the Study of Reading.

Resnick, L. B. (1987). *Education and learning to think*. Washington, D.C.: National Academy Press.

Schuman, H. (1966). The random probe: A technique for evaluating the validity of closed questions. *American Sociological Review, 31*, 218-222.

Spiro, R. J., Vispoel, W. L., Schmitz, J. G., Samarapungavan, A., & Boerger, A. E. (1987). Knowledge acquisition for application: Cognitive flexibility and transfer in complex content domains. In B. C. Britton & S. Glynn (Eds.), *Executive control processes* (pp. 177-200). Hillsdale, NJ: Erlbaum.

Watson, G., & Glaser, E. M. (1980). *Watson-Glaser critical thinking appraisal*. Cleveland, OH: The Psychological Corporation.

Wells, G. L., Ferguson, T. J., & Lindsay, R. C. L. (1981). The tractability of eyewitness confidence and its implications for triers of fact. *Journal of Applied Psychology, 66*, 688-696.

Whimbey, A. (1985). You don't need a special "reasoning" test to implement and evaluate reasoning training. *Educational Leadership, 43*(2), 37-39.

Yarmey, A. D. (1979). *The psychology of eyewitness testimony*. New York: The Free Press.

Table 1

Principles for Appraising Observations

I. Observation statements tend to be more credible than inferences based upon them.

II. An observation statement tends to be credible to the extent that the observer:

1. is alert to the situation and gives his or her statement careful consideration;
2. has no conflict of interest;
3. is skilled at observing the sort of thing observed.

III. An observation statement tends to be credible to the extent that the observation conditions:

1. provide a satisfactory medium of observation;
2. provide sufficient time for observation;
3. include adequate instrumentation, if instrument is used.

IV. An observation statement tends to be credible to the extent that the observation statement:

1. is corroborated;
2. is made close to the time of observing;
3. is not given in response to a leading question.

Table 2

Rating Scale for Assigning Thinking Scores

Rating	Basis of Evaluation
1.....	The examinee indicates a criterion by which a correct comparison of the underlined statements might be made, or the examinee uses a criterion in comparing the two statements but does not explicitly indicate what the criterion is.
2.....	The examinee indicates a criterion and uses it correctly to compare the underlined statements.
3.....	The examinee indicates a criterion, uses it correctly to compare the underlined statements, and justifies the comparison on the basis of a general principle.
0.....	The examinee does none of the above, does not think critically in some other way, or does not respond.

Table 3

Assigning Thinking Scores for Item 3

1.....	The examinee points out that Mr. Wang was involved in the accident.
2.....	The examinee points out that Mr. Wang was involved in the accident and compares Mr. Wang's involvement to Ms. Vernon's being a bystander.
3.....	The examinee points out that Mr. Wang was involved in the accident, compares this with Ms. Vernon's non-involvement, and shows that these facts put Wang in a position of people who stand to profit or lose depending upon what they say.
0.....	The examinee does none of the above, does not think critically in some other way, or does not respond.

Table 4

Two Students' Performance and Thinking Scores for Item 3

Student	Performance Score	Thinking Score
A	1	0
B	0	0

Table 5

Evidence Weightings for Thinking Score/Performance Score Combinations

	Thinking Scores		
Performance Scores	0	1	2
0	+2	-1	-2
1	-1	+1	+2

Table 6

T/P Indexes and Item/Test Biserials for Versions B and C

	T/P Indexes		Item/Test Biserials	
Item No.	Version B	Version C	Version B	Version C
1	.462	.885	.131	.294
2	.731	.885	.185	.323
3	.616	.412	.354	.337
4	.615	.346	.301	.249
5	.885	.731	.186	.409
6	.769	.615	.296	.361
7	.577	.308	.182	.156
8	.385	.385	.437	.258
9	.769	.538	.150	.150
10	.346	.423	.417	.182
11	-.231	.346	.074	.223
12	.583	.387	.201	.302
13	.458	.654	.282	.139
14	.208	.423	.374	.371
15	.625	.692	.123	.219
16	.536	.625	.313	.471
17	.715	.833	.409	.457
18	.964	.958	.447	.454
19	.786	.917	.411	.497

Table 6 (continued)

Item No.	Version B	Version C	Version B	Version C
20	.250	.958	.327	.234
21	.250	.917	.417	.368
22	.429	.750	.299	.221
23	.429	.833	.419	.313
24	.679	.500	.290	.314
25	.769	.958	.488	.374
26	.333	.667	.538	.294
27	.413	.708	.292	.323
28	.750	.708	.365	.337
29	.875	.850	.067	.249
30	.423	.800	.248	.409
31	.346	.800	.349	.361
32	.692	.700	.358	.156
33	.654	.300	.222	.258
34	.538	.750	.462	.150
35	.769	.500	.428	.182
36	.461	.450	.403	.223
37	.731	.800	.520	.302
38	.334	.400	.211	.139
39	.833	.550	.350	.371
40	.625	.550	.428	.219
41	.250	.650	.333	.471
42	.375	.250	.459	.457

Table 6 (continued)

Item No.	Version B	Version C	Version B	Version C
43	.209	.650	.433	.454
44	.875	.700	.264	.497
45	.667	.500	.400	.234
46	.958	.750	.501	.368
47	.833	.950	.476	.221
48	.917	.800	.448	.313
49	.708	.750	.475	.314
50	.750	.950	.329	.374

Table 7

KR-20 Reliability Estimates for TAO, CCTT, and W-G

	Test		
Group	TAO	CCTT	W-G
A	.76	.74	.92
B	.68	.71	.72
C	.73	.78	.76
D	.58	.71	.74
Average	.69	.72	.80

Table 8

Correlations between TAO, CCTT, and W-G

	Test	
Group	CCTT	W-G
A	.62	.37
B	.49	.11
C	.45	.37
D	.35	.41

Appendix A

Interviewing Model for Test on Appraising Observations

Stage One

I.1 Inform examinee of purpose of interview: to find out what he or she is thinking while choosing answers to questions on the test.

I.2 Inform examinee of his or her role: to respond as completely as possible to the questions asked on the test.

Stage Two

II.1 Interviewer says to examinee:
"As you do each question tell me all you can about what you are thinking while you are picking your answers."

II.2 Interviewer can interrupt examinee's narrative only to:

II.2.1 probe for ambiguous reference of demonstratives or third person pronouns by saying:
"Could you tell me what you mean by . . . ?
Example: When doing Item 3 the examinee says: "This driver here is more believable."
Probe immediately: "Could you tell me what you mean by 'this driver here'?"

II.2.2 probe for obvious reading mistakes by saying:
"Did you read . . . ?" (Do not endorse answers.)
Example: When doing Item 3 the examinee reads: "Mr. Wang did use his signal" for "Mr. Wang did not use his signal." Probe immediately: "Did you read 'Mr. Wang *did* use his signal'?"

II.3 Interviewer can respond to examinee's questions only as follows:

II.3.1 If examinee probes for facts, say: "You can only go by what is written."
Example: When doing Item 3 the examinee asks: "Was Vernon close enough to see Wang's car?" Answer only: "You can only go by what is written."

II.3.2 If the examinee probes for reasons, say: "You can decide only according to what is said and what you know."
Example: When doing Item 3 the examinee asks: "How did Vernon know what caused the accident?" Answer only: "You can decide only according to what is said and what you know."

II.4 General cautions:

II.4.1 Do not begin to speak immediately after the examinee stops; give the examinee a few seconds to continue.

II.4.2 Do not cut off examinee's reasoning by in any way signalling that enough has been said, even though the examinee might seek such signals.

II.4.3 Do not endorse or criticize the examinee's fact-finding or reason-giving.

Stage Three

III.1 If "neither" is the answer chosen, then say: "So you believe neither is more believable?" Wait for examinee to respond and *then proceed to next item.*

III.2 If the item is an inference vs. observation question, that is, a question testing Principle I in Table 1, and

III.2.1 If the examinee identifies a criterion[a], makes a comparison on the basis of that criterion, and cites a general principle, *then proceed to the next item.*

III.2.2 If the examinee identifies a criterion, makes a comparison on the basis of that criterion, but cites no general principle, then probe: "So (state the criterion mentioned) makes the difference?" with emphasis on the statement of the criterion. Wait for the examinee to respond and *then proceed to the next item.*

III.2.3 If the examinee identifies a criterion, but makes no comparison on the basis of that criterion, then probe: "Could you tell me more about the difference (state the criterion mentioned) makes to your thinking?" Wait for the examinee to respond and *then proceed to the next item.*

III.2.4 If the examinee does not identify a criterion, then probe: "Could you explain a little more what makes you believe one report more than the other? Wait for the examinee to respond and *then proceed to the next item.*

III.3 If the item tests any principle other than Principle I, and

III.3.1 Same as III.2.1
III.3.2 Same as III.2.2
III.3.3 Same as III.2.3
III.3.4 If the examinee does not identify a criterion, then probe: "Did (state the criterion assumed when designing the item) play any part in your thinking?" If the response is affirmative, then probe: "Could you explain the part it played?" Wait for the examinee to respond and *then proceed to the next item.* If the response is negative, *then proceed to the next item.*

III.3.5 If the examinee explicitly rejects as a criterion the one assumed when designing the item, then probe: "Could you tell me some more about why (state the criterion assumed when designing the item) does not make a difference to your thinking?" Wait for the examinee to respond and *then proceed to the next item.*

III.4 Interviewing strategies for Stage Three

III.4.1 If the examinee identifies more than one criterion including the one assumed when designing the test, then probe about the assumed criterion.

III.4.2 If the examinee identifies more than one criterion not including the one assumed when designing the test, then probe for the first one identified.

III.4.3 If there is doubt about categorizing a response, then choose the less leading probe, that is, the one that comes first on the list.

---

[a]A criterion is a fact or combination of facts cited in the text that can be used to make a comparison between the observers, observation conditions, or observation reports. In Item 3, the criterion is the combination of facts that Wang was involved in the accident and Vernon was not involved.